Report on MovieLens

Raphael Kummer

2023-01-14

MovieLens

Introduction

This is a report on the MovieLens data analysis and a recommendation model training and its performance. First the dataset is to be explored and inspected to evaluate possible training approaches. Next part is building a ML model to recommend movies to users.

Dataset

Grouplens created a movie rating dataset. The 10M dataset (Harper and Konstan 2015) used in this project is a subset of 10 million ratings of 10'000 movies by 72'000 random selected users.

Goal

Given is the loading of the MovieLens 10M dataset, split into an *edx* and a *final_holdout_test* set containing 10% of the MovieLens data. The dataset contains userId, movieId, rating, timestamp, title, and genre.

```
## Loading required package: tidyverse
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## Loading required package: caret
##
## Loading required package: lattice
##
##
## Attaching package: 'caret'
##
##
## The following object is masked from 'package:purrr':
##
##     lift
##
##
## Joining, by = c("userId", "movieId", "rating", "timestamp", "title", "genres")
```

Data Inspection and preprocessing

```
## Loading required package: devtools
```

```
## Loading required package: usethis
```

```
## Loading required package: benchmarkme
```

The *edx* dataset look like this:

	userId	movieId	rating	timestamp	title	genres
1	1	122	5	838985046	Boomerang (1992)	Comedy|Romance
2	1	185	5	838983525	Net, The (1995)	Action|Crime|Thriller
4	1	292	5	838983421	Outbreak (1995)	Action|Drama|Sci-Fi|Thriller
5	1	316	5	838983392	Stargate (1994)	Action|Adventure|Sci-Fi
6	1	329	5	838983392	Star Trek: Generations (1994)	Action|Adventure|Drama|Sci-Fi
7	1	355	5	838984474	Flintstones, The (1994)	Children|Comedy|Fantasy

There are several aspects of the edx dataset to consider exploring: - user ratings in relation to genre - user ratings in relation to movie release year - user ratings in relation to popularity of movies (indie vs blockbuster) - . . .

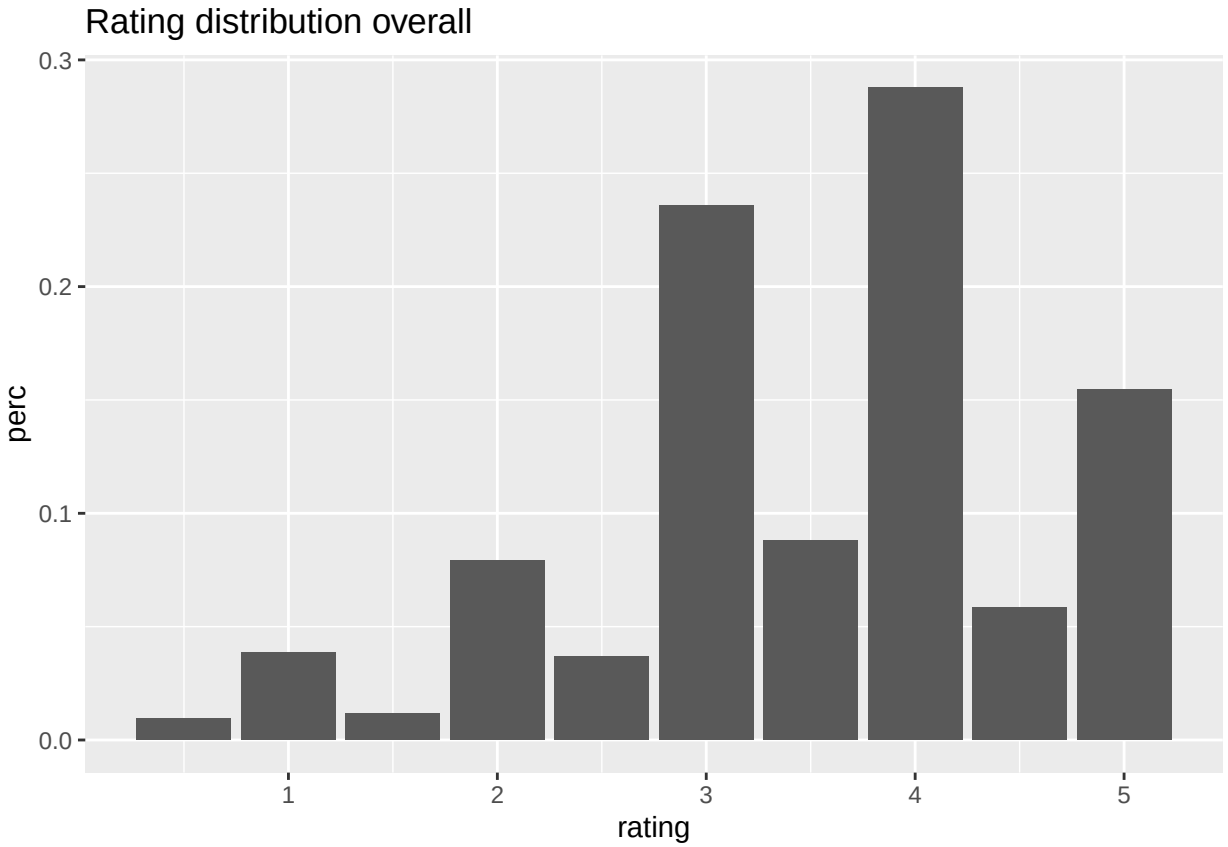
Data cleanup

Check if there are any NA in the dataset.

```
anyNA(edx)
```

```
## [1] FALSE
```

There are no missing values in *edx*.



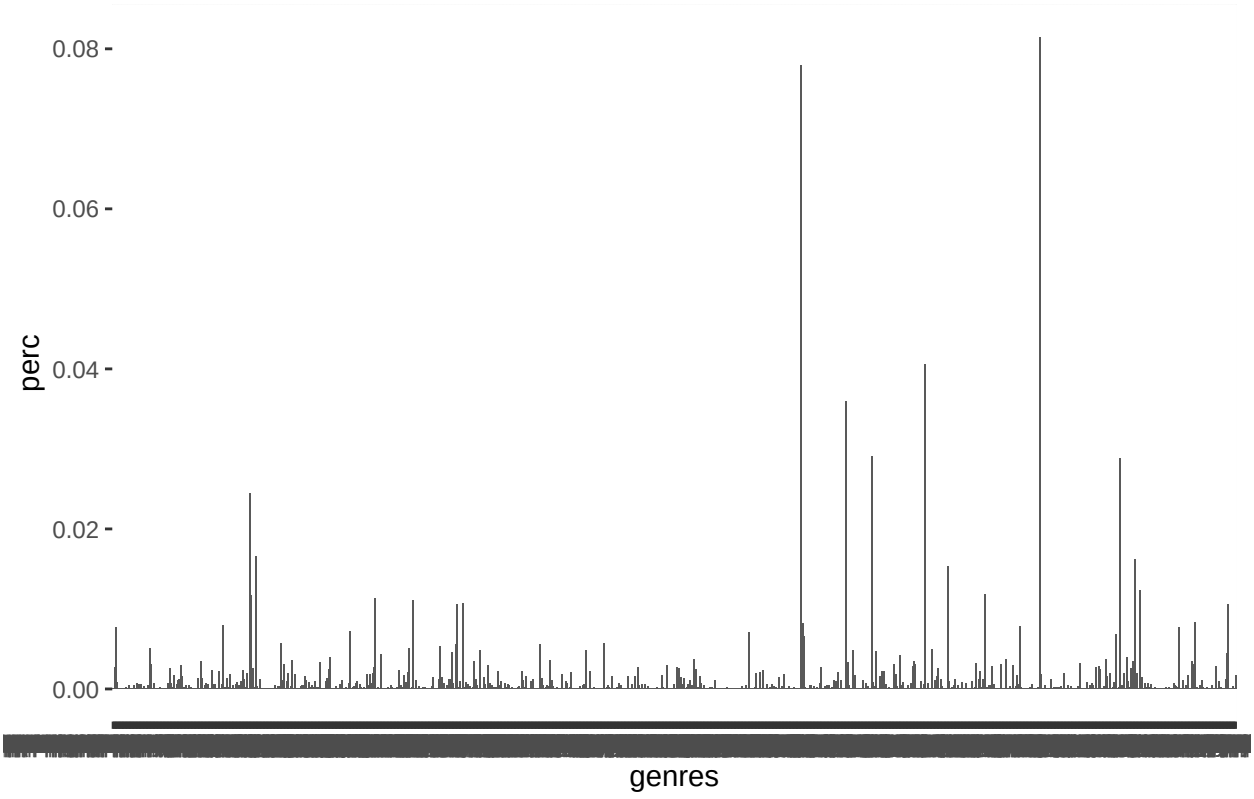
Is the rating of movies dependent of release year of the movie?

```
#mutate()
```

Is the rating dependant of genre? (Of a user. e.g. UserX has 90% of rated movies in the genre of Comedy, he is more likely to rate a Comedy better than Action or Crime)

```
#edx_genres <- edx %>% mutate(comedy = )

edx %>%
  group_by(genres) %>%
  summarise(perc = n()/nrow(edx)) %>%
    ggplot(aes(genres, perc)) +
    geom_col() +
    labs(title = "Genres distribution overall")
```



- check user ratings vs different genres
- ...

Model

Results

RMSE

Conclusion

- summary
- limitations

Future Improvements

- future work

System

Hardware

```
get_cpu()
```

```
## $vendor_id
## [1] "GenuineIntel"
##
## $model_name
```

```
## [1] "Intel(R) Core(TM) i7-7500U CPU @ 2.70GHz"
##
## $no_of_cores
## [1] 4
```

```
get_ram()
```

```
## 8.04 GB
```

All above computations are done with an Intel .... cores.and ... GB RAM. ### Software

```
sessionInfo()
```

```
## R version 4.2.2 (2022-10-31)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Arch Linux
##
## Matrix products: default
## BLAS:   /usr/lib/libblas.so.3.11.0
## LAPACK: /usr/lib/liblapack.so.3.11.0
##
## Random number generation:
##  RNG:     Mersenne-Twister
##  Normal:  Inversion
##  Sample:  Rounding
##
## locale:
##  [1] LC_CTYPE=en_GB.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_GB.UTF-8        LC_COLLATE=en_GB.UTF-8
##  [5] LC_MONETARY=en_GB.UTF-8    LC_MESSAGES=en_GB.UTF-8
##  [7] LC_PAPER=en_GB.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_GB.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] benchmarkme_1.0.8 devtools_2.4.5    usethis_2.1.6     caret_6.0-93
##  [5] lattice_0.20-45   forcats_0.5.2     stringr_1.4.1     dplyr_1.0.10
##  [9] purrr_0.3.5       readr_2.1.3       tidyr_1.2.1       tibble_3.1.8
## [13] ggplot2_3.4.0     tidyverse_1.3.2
##
## loaded via a namespace (and not attached):
##   [1] googledrive_2.0.0     colorspace_2.0-3      ellipsis_0.3.2
##   [4] class_7.3-20          fs_1.5.2              rstudioapi_0.14
##   [7] farver_2.1.1          listenv_0.8.0         remotes_2.4.2
##  [10] bit64_4.0.5           prodlim_2019.11.13    fansi_1.0.3
##  [13] lubridate_1.9.0       xml2_1.3.3            codetools_0.2-18
##  [16] splines_4.2.2         doParallel_1.0.17     cachem_1.0.6
##  [19] knitr_1.41            pkgload_1.3.2         jsonlite_1.8.3
##  [22] pROC_1.18.0           broom_1.0.1           dbplyr_2.2.1
##  [25] shiny_1.7.4           compiler_4.2.2        httr_1.4.4
##  [28] backports_1.4.1       assertthat_0.2.1      Matrix_1.5-1
##  [31] fastmap_1.1.0         gargle_1.2.1          cli_3.4.1
##  [34] later_1.3.0           prettyunits_1.1.1     htmltools_0.5.4
```

```
##  [37] tools_4.2.2           gtable_0.3.1          glue_1.6.2
##  [40] reshape2_1.4.4        Rcpp_1.0.9            cellranger_1.1.0
##  [43] vctrs_0.5.1           nlme_3.1-160          iterators_1.0.14
##  [46] timeDate_4021.106     gower_1.0.0           xfun_0.35
##  [49] globals_0.16.2        ps_1.7.2              rvest_1.0.3
##  [52] timechange_0.1.1      mime_0.12             miniUI_0.1.1.1
##  [55] lifecycle_1.0.3       googlesheets4_1.0.1   future_1.29.0
##  [58] MASS_7.3-58.1         scales_1.2.1          ipred_0.9-13
##  [61] vroom_1.6.0           hms_1.1.2             promises_1.2.0.1
##  [64] parallel_4.2.2        yaml_2.3.6            memoise_2.0.1
##  [67] rpart_4.1.19          stringi_1.7.8         highr_0.9
##  [70] foreach_1.5.2         hardhat_1.2.0         pkgbuild_1.4.0
##  [73] benchmarkmeData_1.0.4 lava_1.7.0            rlang_1.0.6
##  [76] pkgconfig_2.0.3       evaluate_0.18         labeling_0.4.2
##  [79] htmlwidgets_1.6.1     recipes_1.0.3         bit_4.0.5
##  [82] tidyselect_1.2.0      processx_3.8.0        parallelly_1.32.1
##  [85] plyr_1.8.8            magrittr_2.0.3        R6_2.5.1
##  [88] profvis_0.3.7         generics_0.1.3        DBI_1.1.3
##  [91] pillar_1.8.1          haven_2.5.1           withr_2.5.0
##  [94] survival_3.4-0        nnet_7.3-18           future.apply_1.10.0
##  [97] modelr_0.1.10         crayon_1.5.2          utf8_1.2.2
## [100] urlchecker_1.0.1      tzdb_0.3.0            rmarkdown_2.18
## [103] grid_4.2.2            readxl_1.4.1          data.table_1.14.6
## [106] callr_3.7.3           ModelMetrics_1.2.2.2  reprex_2.0.2
## [109] digest_0.6.30         xtable_1.8-4          httpuv_1.6.8
## [112] stats4_4.2.2          munsell_0.5.0         sessioninfo_1.2.2
```

Resources

[1] Rafael Irizarry. 2018. Introduction to Data Science. https://rafalab.dfci.harvard.edu/dsbook/

Harper, F. Maxwell, and Joseph A. Konstan. 2015. "Thje MovieLens Datasets." *ACM Transactions on Interactive Intelligent Systems*, no. 4 (December): 1–19. https://doi.org/10.1145/2827872.